Describing the quality of research datasets across disciplines: a comparative study

Tiffany C. Chao *tchao @ illinois.edu*

Center for Informatics Research in Science and Scholarship

Graduate School of Library and Information Science, University of Illinois

OVERVIEW

Descriptions of data quality play an integral role in data reuse determinations by potential users. However, *quality* characterizations vary across different research cultures where formalized criteria may be limited or may not be sufficient for curation purposes.[1]

Addressed in this poster is a preliminary examination of research dataset records from three sub-disciplines of Earth science that centers on:

- how *quality* is described by scientists for research datasets
- what patterns in quality description emerge across different sub-disciplinary fields

ASSESSMENT OF QUALITY DESCRIPTIONS

SAMPLE SCOPE: (3) sub-disciplines of Earth science from the Global Change Master Directory (GCMD) (http://gcmd.nasa.gov/)

- Geochemistry
- Population science
- Atmospheric science

SOURCE: Dataset metadata records from GCMD (collected in Fall 2012); records follow the DIF format (Directory Interchange format) which includes a specific field for *quality.*

APPROACH: Extracted available descriptions from the <quality> field for each dataset record and assigned a category based on prescribed DIF "quality" definitions. Reviewed initial categories for emergent topics and recoded descriptions with new list. Only one category was assigned to each description.

<quality> definitions for DIF records

The <quality> field allows the author to provide information about the quality of the data or any quality assurance procedures followed in producing the data described in the metadata.[2]

This information may include:

- Indicators of data quality or quality flags
- Established quality control mechanisms
- Established quantitative quality measurements
- Recognized or potential problems with data quality

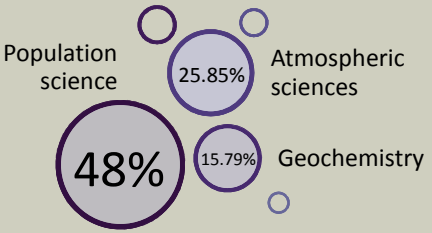
The <quality> field is highly recommended but not required for the DIF record

Observed <quality> categories in DIF records across all three sub-discipline areas

- **Acknowledgements:** for funding or external support
- **Data description:** details about the provided data (i.e. format, variable names, study context, etc.)
- **Disclaimer:** warning about using the data (i.e. data provider not liable for accuracy)
- **Method:** details about how the data were collected and analyzed
- **Problem**: statements related to the completeness of data provided (i.e. gaps in data collection, potential data corruption, etc.)
- **Quality assurance procedures:** identified persons or procedures used for quality checks and review of data
- **Quality indicator:** usefulness of provided data; marked status of provided data (i.e. "as is", "approximate values)
- **Reference:** to an external source of information (i.e. data provider contact information; physical location of data; URL, etc.)

FINDINGS

What percentage of available DIF records have information about quality?



Similarities in quality descriptions are observed across all three sub-disciplines with the categories of *Reference* and *Method*. With each, the inclusion of detailed information about data collection techniques and methodological processes complements findings from qualitative studies on scientists' assessment criteria of data for reuse. [3,4]

The *Problems* related to provided data varied in the level of detail describing the nature of the issue. For instance, some statements about the atmospheric science data discussed specific areas where the data provided may be compromised or unreliable whereas other statements were more general.

These described problems were also examined in relation to data types and whether similar issues were identified within each of the sub-discipline fields. Though three data types, or formats, were found in common (ASCII, PDF, and shapefiles), descriptions of quality did not discuss problems but spanned the other observed quality categories.

The top three observed quality categories for datasets in each sub-discipline along with examples from the record descriptions are listed below. The number after each category indicates the percentage of records that fall under that particular category.

Geochemistry

- **Reference (31.37%)**
  - external information source (i.e. publications, digital data repository, physical location of samples, investigator contact information)
- **Quality indicator (29.41%)**
  - data values are "approximate"; "good, high precision"
- **Method (17.65%)**
  - location of data collection, time period for collection, analyses procedures detailed

Population Science

- **Problem (33.33%)**
  - Quality information "not available" or "unknown"
- **Disclaimer (16.67%)**
  - no "warranties" or "guarantees" that data provided are accurate or appropriate for user purposes
- **Method/Data description/Reference (12.5% each)**
  - analytic procedures named, "standard scientific procedures"; context information about data collection or data source, sample size detailed; URL to external site

Atmospheric Science

- **Reference (27.61%)**
  - contact investigators or data providers about quality and limitations of the data; review URL for more information
- **Problem (21.20%)**
  - "data are not quality checked"; anomalies present; temporal gaps in data coverage
- **Method (15.95%)**
  - data collection procedures (i.e. processing, calibration, instruments used, analysis); issues with data collection

CONCLUSIONS

The quality of data can be described in a number of ways. In applying the DIF field definitions for quality to available records, different interpretations emerged resulting in additional categories. To a certain extent, some of the observed categories may actually fit in with existing definitions while others may not be related to quality at all but to other aspects of the data.

The comparison of data quality descriptions between geochemistry, population science, and atmospheric science revealed similar types of information publically conveyed, with a focus on how the data were generated (Method) and what additional evidence was available (Reference) to potentially establish quality. The amount of detail presented did vary within each sub-discipline area where no consistent patterns were observed.

Directions for future work consider how informative current data quality descriptions actually are as assessed by research scholars and curation professionals and how these descriptions could be improved. Such assessments have implications for the level of curatorial service allocated to the dataset in order to support future access and reuse.

REFERENCES:

[1]Marchionini, G., Lee, C., Bowden, H., & Lesk, M. (2012). Curating for quality ensuring data quality to enable new science. Final report: invitational workshop sponsored by the National Science Foundation. Retrieved from http://datacuration.web.unc.edu/

[2]"Quality", Directory Interchange Format (DIF) Writer's Guide. (2012). Global Change Master Directory. *National Aeronautics and Space Administration*. Retrieved from http://gcmd.nasa.gov/User/difguide/quality.html

[3]Zimmerman, A. (2007). Not by metadata alone: the use of diverse forms of knowledge to locate data for reuse. *International Journal on Digital Libraries*, 7(1), 5–16.

[4]Faniel, I. M., & Jacobsen, T. E. (2010). Reusing scientific data: How earthquake engineering researchers assess the reusability of colleagues' data. *Computer Supported Cooperative Work (CSCW)*, 19(3-4), 355–375. doi:10.1007/s10606-010-9117-8